Searching for connections in journalistic data lakes

Ioana Manolescu
Inria Saclay and Ecole Polytechnique, Palaiseau, France

1 Context

Digital data, whether text (news articles), semi-structured (tweets, other social media content) or structured (RDF or CSV files) is being produced and shared at very large speed today. Events such as elections, public manifestations, disclosures of illegal or surprising activity etc. unfold under the form of new data items being created and added to the global corpus of available information. Integrating and making sense of this wealth of data is a highly prized goal.

We consider the particular setting of a data lake, which is a set of heterogeneous data sources, in which some common elements may be found (e.g., people, places, organization names, or semantic concepts). For instance, a media article may mention a politician and a law she contributed to, a relational database of companies may list their owners and main officers, finally a Wikipedia page may connect the politician to her spouse that is a CEO of a company concerned by the law.

We consider a graph view of the data lake, reflecting all the connections between the entities and objects mentioned in any of the data sources. For instance, we may want to answer questions such as: “What are all the connections between a politician named Smith and a company in the security sector”? or: “What are the links between Mr. Jones, company ABC and company XYZ”? An answer to such a query is a tree, whose leaves contain or match the search conditions specified by the user; the useful information in an answer resides in the connections between them. As there may be a large number of answers, an score can be assigned to each answer; the problem, then, becomes to compute the k highest-score answer trees for a given query, as indicated by the score function. A first prototype, called ConnectionLens [CDG+18], has been developed in collaboration by the CEDAR team, the AIST institute of Japan and the University of Lisbon, Portugal to compute such answers, and work on the prototype is still ongoing.

2 Internship goal

The goal of the internship is to improve ConnectionLens search performance, investigating one or both of the following directions:

Improving search speed As stated above, the ConnectionLens search problem is very close to the search for the minimum-cost tree connecting a set of nodes in a weighted graph (the so-called Steiner Tree Problem, or STP in short), known to be NP-hard. Thus, efficient algorithms need to make use of heuristics and/or pre-compute (or re-use) partial search results to improve the performance. Parallelism (multi-threading on a single machine, or through implementation on a massively distributed platform) could also be used.

Improving result quality Beyond the classical result score metrics introduced in Information Retrieval for keyword search on text documents, a large variety of tree scoring functions have been proposed to rank the results of keyword search in XML documents, relational databases [HP02, LOF+08], and RDF graphs [LLKD14]. Finding a score function (or functions) most adequate or interesting in our heterogeneous graph context, though, remains an open problem.

Work on ConnectionLens is part the ANR ContentCheck project (http://contentcheck.inria.fr/), where we collaborate with fact-checking journalists from Le Monde (https://www.lemonde.fr/les-decodeurs/) to devise data management methods for data journalism and journalistic fact-checking. We also expect to continue collaborating with our foreign partners from Japan and Portugal.

1https://en.wikipedia.org/wiki/Steiner_tree_problem
3 Required skills and competencies
We welcome candidates with a strong background in algorithms, programming, good communication skills in English (preferred) or French, and a strong scientific curiosity. A good background in data management (databases, distributed computing) would be a plus.

4 Practical information

Location The internship will be carried out in our team, in Palaiseau, France (https://team.inria.fr/cedar/contact/)

Remuneration Approx. 500 €/month (more can be discussed for M2 students)

Contact Ioana Manolescu (ioana.manolescu@inria.fr), http://pages.saclay.inria.fr/ioana.manolescu/

How to apply Please send to ioana.manolescu@inria.fr an email with a short CV and grade transcript in the relevant courses.

References


[HP02] Vagelis Hristidis and Yannis Papakonstantinou. DISCOVER: keyword search in relational databases. In VLDB, 2002. 1

[LLKD14] Wangchao Le, Feifei Li, Anastasios Kementsietsidis, and Songyun Duan. Scalable keyword search on large RDF data. TKDE, 26(11), 2014. 1